UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 – Innføring i anvendt statistikk
Eksamensdag:	Tirsdag 12. desember 2017
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 5 sider	
Tillatte hjelpemidler:	Godkjent kalkulator, ordliste for STK1000, og lærebok (alle utgaver, og det er lov å notere i læreboka)

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.
Alle deloppgaver teller likt i vurderingen av besvarelsen. Lykke til!

Oppgave 1

John Arbuthnot blir av mange tildelt æren for å ha beregnet verdens første p-verdi. Nullhypotesen hans om at det blir født like mange gutter som jenter ble undersøkt ved å studere dåpsstatistikken i London mellom 1629 og 1710. Han observerte at i 82 år på rad ble det født (og døpt) flere gutter enn jenter. I denne oppgaven skal vi se nærmere på om det var nødvendig med observasjoner over 82 år, eller om John Arbuthnot kunne klart seg med ett år.

Chriſtened.

Anno.	Males.	Females.
1629	5218	4683
30	4858	4457
31	4422	4102
32	4994	4590
33	5158	4839
34	5035	4820
35	5106	4928
36	4917	4605
37	4703	4457
38	5359	4952
39	5366	4784
40	5518	5332
41	5470	5200
42	5460	4910
43	4793	4617
44	4107	3997
45	4047	3919
46	3768	3395
47	3796	3536

Chriſtened.

Anno.	Males.	Females.
1648	3363	3181
49	3079	2746
50	2890	2722
51	3231	2840
52	3220	2908
53	3196	2959
54	3441	3179
55	3655	3349
56	3668	3382
57	3396	3289
58	3157	3013
59	3209	2781
60	3724	3247
61	4748	4107
62	5216	4803
63	5411	4881
64	6041	5681
65	5114	4858
66	4678	4319

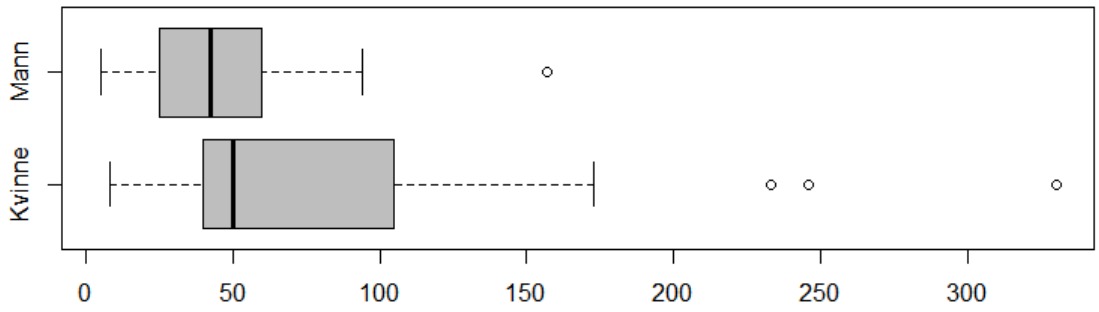
a) Beregn sannsynligheten for å observere flere guttefødsler enn jentefødsler i 82 år på rad, dersom sannsynligheten for å føde en gutt er lik sannsynligheten for å føde en jente.
La den stokastiske variabelen X være antall guttefødsler i 1629. Hvilken sannsynlighetsfordeling har X?
Hvilke tre forutsetninger må være oppfylt for at X skal ha denne fordelingen?

b) Ta utgangspunkt i X og parameteren p i fordelingen fra oppgave a). Formuler nullhypotesen og den alternative hypotesen for hvor stor andel gutter som blir født i 1629.
Gi et punktestimat for p, og beregn et 95% konfidensintervall for p.
Gi en tolkning av konfidensintervallet, og forklar hvilken konklusjon du kommer frem til på hypotesetesten hvis du bruker et signifikansnivå på 0.05.

c) Anta at den sanne andelen guttefødsler er 0.52. Hvor mange fødsler (n) måtte du observere for å få forkastet en tosidig H_0 på nivå 0.05? (Hint: Ta utgangspunkt i nedre grense i et 95% konfidensintervall, som i oppgave b).)

Oppgave 2

Da STK1000-studenter høsten 2016 ble spurt om hvor mange hverdagsklær de har i skapet sitt, fordelte antallet klær seg på 20 menn og 31 kvinner på følgende måte:



Deskriptiv statistikk for de 20 mennene:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      sd
    5.0    27.5    42.5    49.0    59.0   157.0    34.3
```

Deskriptiv statistikk for de 31 kvinnene:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      sd
    8.0    40.0    50.0    85.0   105.0   330.0    75.2
```

Utskriftene viser resultatene fra en to-utvalgs t-test og en Wilcoxon Rank Sum test:

```
> t.test(hverdag~kjonn)

        Welch Two Sample t-test

data:  hverdag by kjonn
t = 2.3216, df = 45.082, p-value = 0.02483
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.780768 67.383748
sample estimates:
mean in group Kvinne   mean in group Mann
            85.03226             48.95000

> wilcox.test(hverdag~kjonn)

        Wilcoxon rank sum test with continuity correction

data:  hverdag by kjonn
W = 399.5, p-value = 0.08544
alternative hypothesis: true location shift is not equal to 0
```

a) Hvilke oppsummeringstall fra den deskriptive statistikken beskriver gruppene best? Begrunn svaret.

b) Sett opp hypoteser for å teste om det er forskjell på antall klær mellom kjønnene. Forklar hvilken test du velger og hvorfor.
Velg signifikansnivå 0.05. Hvilken konklusjon trekker du og hvorfor?

Oppgave 3

Blant ikke-gravide voksne er det funnet en sammenheng mellom høy body mass index (BMI), altså (vekt i kg)/(høyde i m)^2, og høy insulinresistens. Insulinresistensen reflekteres i høye blodsukkerverdier, spesielt etter at man har spist. I en studie av 130 gravide kvinner ønsket man å undersøke sammenhengen mellom kvinnenes BMI før graviditeten, og deres blodsukkernivåer (målt i mmol/l) to timer etter matinntak, i siste halvdel av svangerskapet. Følgende regresjonsanalyse ble gjort:

```
> summary(lm(blodsukker ~ BMI ))

Call:
lm(formula = blodsukker  ~ BMI )

Residuals:
    Min      1Q  Median      3Q     Max
-2.8228 -0.7290 -0.2273  0.7749  3.5811

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77942    0.88482   3.141  0.00209 **
BMI          0.08225    0.03897   2.111  0.03676 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.169 on 128 degrees of freedom
Multiple R-squared:  0.03363,   Adjusted R-squared:  0.02608
F-statistic: 4.454 on 1 and 128 DF,  p-value: 0.03676
```

a) Hva er effektmålet her?
Hvordan tolkes det?
Gi et estimat for sammenhengen mellom mors BMI før svangerskapet og hennes blodsukkernivå to timer etter matinntak, i siste halvdel av svangerskapet.
Beregn et 95% konfidensintervall for denne sammenhengen.

Forskerne mistenkte at vektøkningen i svangerskapet kunne være en viktig faktor i dette. Det ble lagt til grunn forskning fra ikke-gravide, som viste at selv en moderat økning i BMI førte til økt insulinresistens hos personer som ble fulgt opp over tid.
Helseråd som gis til gravide om vektøkning i svangerskapet har som mål å redusere uheldige konsekvenser av overvekt, både for den gravide og barnet hun bærer. Anbefalt vektøkning vil derfor være en konsekvens av kvinnenes BMI før svangerskapet. Overvektige kvinner anbefales en mindre vektoppgang enn normalvektige kvinner.

b) Kan vektøkningen i svangerskapet antas å være en en konfunderende variabel (confounder eller lurking variable) for sammenhengen mellom mors BMI før svangerskapet og hennes blodsukkernivå to timer etter matinntak, sent i svangerskapet? Begrunn svaret.
Kommenter følgende utskrift i lys av det du nettopp svarte.

```
> summary(lm(blodsukker  ~ BMI + vektokning))

Call:
lm(formula = blodsukker  ~ BMI  + vektokning)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8334 -0.6902 -0.2317  0.8012  3.5537

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.63692    0.97384   2.708  0.00771 **
BMI          0.08268    0.03912   2.113  0.03653 *
vektokning   0.01789    0.05023   0.356  0.72231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.173 on 127 degrees of freedom
Multiple R-squared:  0.03459,   Adjusted R-squared:  0.01939
F-statistic: 2.275 on 2 and 127 DF,  p-value: 0.1069
```

Oppgave 4

Et av datasettene man finner i R er et datasett med høyde, omkrets og tømmervolumet i 31 felte kirsebærtrær. Omkretsen til trærne (i cm) er målt i brysthøyde, 137 cm over bakken. I denne oppgaven er tømmervolumet kovertert til liter. Man kan gå ut i fra at en liter tilsvarer en vedkubbe. Dersom man plotter tømmervolumet y_i (kalt `vedkubbe` i utskriften) og omkretsen på treet x_i (kalt `omkrets` i utskriften) i et scatterplot, ser man at den kan uttrykkes ved regresjonsmodellen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0,\sigma).$$

Gjennomsnitt og standardavvik er $\bar{x} = 105.7$, $sd_x = 25.0$, og $\bar{y} = 854.3$, $sd_y = 465.5$.
Utskriften viser en regresjonsanalyse som ble gjort av disse 31 trærne:

```
> summary(lm(vedkubber~omkrets))

Call:
lm(formula = vedkubber ~ omkrets)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-228.386  -87.972    4.303   98.961  271.468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1046.1223    95.2903  -10.98 7.62e-12 ***
omkrets        17.9769     0.8779   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.4 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

a) Hvordan må sammenhengen mellom x og y i scatterplottet se ut for at det skal være fornuftig å bruke denne regresjonsmodellen?

   Gi estimater og tolkning av estimatene for parameterne β_0 og β_1.

b) Beregn et 95% prediksjonsintervall for tømmervolum målt i vedkubber for et tre med en omkrets på 94.2 cm (diameter 30 cm).

c) Utskriften under viser en tilsvarende regresjonsanalyse for trærnes omkrets mot trærnes høyde (målt i meter). Et tilhørende 95% prediksjonsintervall for høyden til et tre med en omkrets på 94.2 cm (diameter 30 cm) er [19.2, 26.2]. Forklar hva et prediksjonsintervall viser.
   Gi en tolkning av de to prediksjonsintervallene fra b) og c).
   For de 31 kirsebærtrærne: Hva gir omkretsen til treet en mest presis prediksjon av: tømmervolumet eller høyden på treet? Begrunn svaret.

```
> round(c(mean(trehoydeimeter), sd(trehoydeimeter)),1)
[1] 23.2  1.9
> summary(lm(trehoydeimeter~omkrets))

Call:
lm(formula = trehoydeimeter ~ omkrets)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8349 -0.8439  0.0964  0.7537  3.0314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.90714    1.33603  14.152 1.49e-14 ***
omkrets      0.04027    0.01231   3.272  0.00276 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.688 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

SLUTT